A METHOD FOR ASSOCIATING PATTERNS OF MOTION IN EVENTS FROM VIDEO: A CASE STUDY

Nikolaos Bourbakis[1,2] *and George Bebis*[3]
[1]Wright State University, OH, [2]AIIS Inc., OH, [3]Univ. of Nevada-Reno, USA

Abstract

Representing, associating, interpreting and learning behavioral patterns and rules from observation is always a challenging research problem for cognitive systems and human computer interaction. This problem becomes more difficult in complex environments with many unexpected behavioral patterns from objects, or cells, or individuals or small group of people acting under certain plans at the same time in different locations. This paper targets the development of a novel synergistic methodology for automatically monitoring, representing,, associating and interpreting behavioral patterns and activities, and learning rules from complex environments using videos.

1. INTRODUCTION

It is well known from scientific experiments that the human ability of intensive and continuous observation of dynamically changed scenery for lengthy periods of time is limited and unreliable. However, always humans are called to perform unsuccessfully these tasks. An example of a complex environment is a multiple cameras surveillance system where a human user is responsible of continuously observing many TV screens at the same time for detecting, extracting, associating changes and activities occurring in different places at the same or concurrent time and for long period of time. Another example of such an environment is the long time continuous and intensive observation of psychologically ill patients by a medical doctor for extracting and associating behavioral patterns to the patient's illness. A similar environment is the continuous and lengthy observation of the activity and behavioral patterns of cell populations by using a microscope. In all these and many similar cases humans are called to perform these tedious and difficult tasks with high risk of errors. On the other hand the latest impressive achievements of the computer based technologies offer a valuable assistance to humans who deal with complex and high risk environments for automatically extracting, recognizing and understanding multiple activity in video.

The problem of modeling,, associating, understanding and interpreting behavioral patterns and rules (activities) and their semantics from observation is very challenging. Analyzing activity for interpretation purpose means parsing temporal sequences of object observations to produce high-level descriptions of agent actions and multi-agent interactions. Several research efforts has been performed lately in this area and several interesting approaches have been developed for interpreting human activities from images, however, the capabilities of these systems are quite limited. It is now becoming more evident that meaningful interpretation of visual behavior can not merely rely on a simple mapping from recognized patterns of motion to semantics. In most cases, the same behavior may have several different meanings depending upon the scene and task context in which it is performed. In addition, in the recent years the scientific accomplishments in image understanding and visual languages fields have been shifted into the video understanding domain. Several research efforts have been presented in the literature to achieve objects motion in a sequence of images, and some cases basic understanding of human behavior, such as description of a human that carries a box, a human that walks, and simple activity description [3-7,19,29-31,33,34,37-39]. Specifically, relevant research work has been done on trajectory guided tracking and recognition of actions, stochastic temporal models of human activities, Bayesian approach to human activity recognition, complex visual activity recognition using a temporally ordered database, layered probabilistic recognition of human action, learning and recognizing human dynamics in video sequences, probabilistic recognition of human actions, recognizing human action in time-sequential images using hidden Markov models, visual understanding of dynamic hand gestures etc. In [29], a method for real-time 3D tracking of human motion through multiple 2D camera views using a discrete relaxation algorithm is described. Adaptive appearance models were used in [12] to track in real time a person's forearms and hands under self-occlusion and interpret associated behavioral events. Principal components analysis is used in [30] to model sets of exemplar activities. Recognition is performed by parameterizing the search in the space of admissible transformations that the activities can undergo. In [15], object trajectories were produced using adaptive background subtraction while unsupervised statistical learning was used then to cluster them into descriptions of normal activity. A system called W4 is described in [16] which is capable of locating people, segmenting their body parts, build simple appearance models for tracking, disambiguate between and separately track multiple people in a group, and detect carried objects such as boxes and backpacks. In [31], a system is described which makes context-based decisions about the actions of people in a room. The system uses prior knowledge of the layout of the room. Action recognition is carried out using a state model.

HMMs have been used extensively for visual gesture recognition. In this case, it is feasible to hand-design an adequate transition topology, which is the dominating constraint in the learning problem. However, their usefulness for more complicated systems, such as recognizing human activities, is serious affected by the fact that for models of nontrivial size, one must seek for an appropriate topology using very expensive search techniques. In any case, HMMs have been a common thread in much of the recent work in action recognition [18,19]. Coupled HMMs were used in [38] to detect and classify interactions consisting of two interleaved agent action streams. To cluster video sequences into events and create classifiers to detect those events, an entropy-based minimization approach was proposed in [37] for estimating HMM topology and parameter values simultaneously.

Gestures and multi-object interactions are recognized in [17] by parsing a stochastic context-free grammar that defines multiple events that can be occurring simultaneously in the scene. In [13], the spatio-temporal patterns of simple human activities (i.e., person walking or running) are modeled by the trajectory of multivariate observation vectors which include features such as position, speed, and shape boundary. Each pattern is represented in a low dimensional space and learning is used to capture the dynamics of the activities that they model. Pixel-energy histories are used in [14] to learn normal and abnormal behavioral events while Belief Nets are used to model the semantics of plausible human body motion configurations. The most successful efforts presented or proposed by researchers are mainly based on Hidden Markov Models. These methods, however, don't integrate and incorporate structural and functional representation of knowledge in the same medium and don't offer timing, parallelism, concurrency and synchronization of events needed by demanding applications, such as surveillance systems using many cameras in crowded environments (airports, arenas, etc).

2. THE METHODOLOGY OF ASSOCIATIONS

This paper addresses the need to track humans and objects in a visual scene, and assess the activities in the scene. Starting from low-level vision operations of segmentation and contour generation, features extraction and characterization of the scene objects as graph descriptions will allow recognition and characterization of the objects. From object characterization, a correspondence from structural object features to functional attributes will result in a description of the activity.

The methodology proposed here is based on the implementation of five main components: (1) image and object of interest segmentation (2) extraction and representation of objects of attention, (3) recognition and classification of targeted objects, (4) tracking and association of extracted targeted objects from extracting behavioral patterns and learning rules, and (5) recognition of objects' activity. Here, by object we mean, humans, animals, cells, etc.

2.1 Region based Segmentation

Segmentation is one of the first important and difficult steps of image analysis and computer vision, and many algorithms have been developed [20-24,35]. Here, we intend to use our Fuzzy-like Reasoning Segmentation (*FRS*) method that adds light model as one of the segmentation factors. Its result is more accurate in terms of perception and more suitable for later reconstructing work. The FRS method has three stages (smoothing, edge detection and segmentation) [9]. The initial smoothing operation is intended to remove noise. The smoother and edge detector algorithms are also included in this processing step [23,24]. The segmentation algorithm uses edge information and the smoothed image to find segments present within the image.

2.2 The L-G Graph

The graph is a very powerful methodologies developed for a great variety of the computer science problems. Relational graphs are considered as a good approach to describe pictures or scenes for pattern recognition [25-29]. Our *Local-Global* (*L-G*) graph method adds local part information into graph [3,10]. The graph is a more accurate representation of an object. Thus, we avoid using a non-linear graph matching function. By combining the *FRS* method and the *L-G* graph method, we can improve object recognition accuracy without increasing computation complexity. It is wise not to say much, since the local and global graphs used here are not new but incremental methodologies for efficiently representing images via their visible features and the features relationships. Thus, here the L-G graph is capable of describing with adjustable accuracy and robustness the features contained in an image. The main components of the L-G graph are: (i) the local graph that represents the information related with shape, size, and (ii) the skeleton graph that provides in formation about the internal shape of each segmented region. The global graph represents the relationships among the segmented regions for the entire image. The nodes Pi of the global graph include the L-G, the color, the texture and the skeleton graph. The local-global image graph components are briefly described below.

The Region or Local Graph **(**G) holds information of a contour –line of an image region after segmentation: $G = N_1 a^c{}_{12} N_2 a^c{}_{23} N_3 \ldots N_k a^c{}_{k1} N_1 \otimes N_i\ a^p{}_{ij} N_j \otimes \ldots \otimes N_n a^{rd}{}_{nm} N_m \ldots$

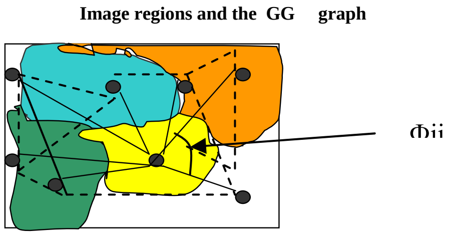
where, $\otimes$ represents the graph relationship operator, and each Ni maintains the structural features of the corresponding line segment, thus, Ni = { sp, orientation (o), length (le), curvature (cu)}, and a_{ij} holds the relationships among these line segments, thus, a_{ij} = { connectivity (c) , parallelism (p) , symmetry (s), relative magnitude (rm), relative distance (rd), etc}.

The Skeleton Graph is also a part of the L-G graph by offering additional information about the regions and is based on the efficient generation of the regions skeletons after the regions segmentation process. The line segments of the skeleton of a region are interrelated with each other through a graph with attributes in a similar way with the one generated by the contour of a segmented region. Thus, the skeleton graph of a region's skeleton is

$$Gsk = K_1 a^c{}_{12} K_2 a^c{}_{23} K_3 \ldots K_k a^c{}_{kq} K_q \otimes K_i\ a^p{}_{ij} K_j \otimes \ldots \otimes K_n a^{rd}{}_{nm} K_{m\ \ldots}$$

where, ⊗ represents the graph relationship operator, and each Ki maintains the structural features of the corresponding line segment, thus, Ki = { sp, orientation (o), length (le), curvature (cu)}, and a_{ij} holds the relationships among these line segments, thus, a_{ij} = { connectivity (c) , parallelism (p) , symmetry (s), relative magnitude (rm), relative distance (rd), etc}. The missing elements for a global visual perception of an image are: the color (or texture) of each region, its relative geographic location (distance and angle) among the other regions, its relative size in regards with the other regions, etc. One way to obtain these additional features is the development of the global image graph (GG).

The Image Global Graph attempts to emulate a human-like understanding by developing global topological relationships among regions and objects. More specifically, for each image region M_i, a skeletonization task is performed and the final centroid GCg(i,x,y) is defined [23].



$$GGA_{(N1)}) = (P_1 R_{12} P_2)\ \Phi_{23}\ (P_1 R_{13} P_3)\ \ \Phi_{34}\ (P_1 R_{14} P_4) \ldots \ldots$$

Fig. 1. The L-G graph of a synthetic image consisted of 9 segmented regions

When all the final centroids have been defined for every image region, the global image graph is developed: $GG(Ak) = (P_1R_{12}P_2)\ \Phi_{23}\ (P_1R_{13}P_3) \ldots (P_1R_{1n-1}P_{n-1})\ \Phi_{n-1n}\ (P_1R_{1n}P_n)$

where P_i is a node that represents a region graph, its color, and its GCg(i,x,y) and the skeleton graph (Gsk), Rij represents the relative distance between two consecutive Gcg, and the orientation of each dg, Φij represents the relative angle between consecutive distances dg(i) and dg(j), see figure 1.. An important feature of the L-G graph is its ability to describe 3-D scenes. The only difference between 2-D from 3-D is that in 3-D the local graph will represent 3-D surfaces and the global graph will appropriately interrelate them .

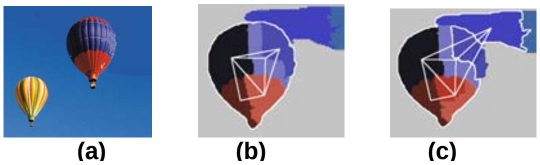


Figure 2 : An example of a mismatch. (a) is the original data image. (b) and (c) show two graphs with one PCRP change. (b) and (c) have a similar graph and also the same region relationships.

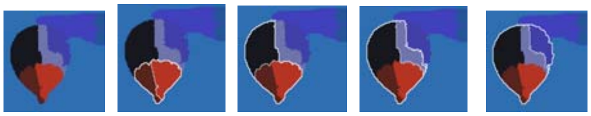


Figure 3. It shows the 5 consecutive steps of the region synthesis for the construction of a balloon.

2.3 Model Driven Region Synthesis/Recognition [11]

For regions and objects recognition, many methods are proposed based on shape similarity, regions relationships or combinations. They assume that the shape has been segmented from the background, and the mathematic shape representation is not sensitive to some kinds of deformations. This is not true in most of the cases, for example, if the shape of a region changes the local graph and the skeleton graph record the changes attached them to the L-G graph. In addition, the centroid may change and the global graph will register that change as well.

The process used here for objects recognition is based on the synthesis of segmented adjacent regions, using the L-G graph, and association (comparison) of the integrated region models available in a database. The method can search object in one given image based upon the provided object model database. In the model database, all objects are represented by their multi-view structure. Every object has been modeled from several (6 max) views. Every view represents a view direction that the user interested. They are not necessary orthogonal views or 3-dimention views. In fact, the views are not required to be orthogonal at all. Any different view of an object, not available in the DB, can be generated as a synthesis of other view from the DB. Figure 2 shows the synthesis of the regions based on the L-G graph. Figure 3 presents the synthesis of the regions selected in 5 steps. Figure 4 shows the structure of the multiple-view, in the L-G database.

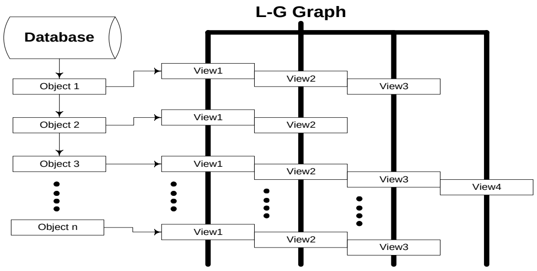


Figure 4: Multi-View Image Database Structure

2.4 Tracking Moving Objects in Sequences of non-consecutive Frames

The tracking of a moving object (or human) in consecutive frames is already an easy task performed by several methodologies available in the literature[4-8,12-19,29-33,34,36]. Here we present a different case where the tracking of an object has to be successful in a number of non-consecutive frames. The reasoning behind this assumption is that there are cases where the moving target goes behind obstacles and its activity has to be continuously recorded and associated with its previous pattern of behavior. In this case our tracking method is based on the synthesis of the regions that compose that target and a continuous comparison with its regions recognized from the latest frame, see figure 5.

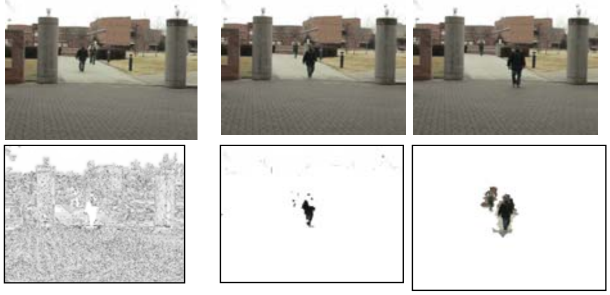


Fig.5: It shows 3 consecutive frames; the spatio -temporal diffusion result and the motion activity map; the segmentation result with kernel density estimation and the segmentation result using kernel density estimation

T [illegible] f the L-G graph. Results of our tracking and recognition method in a natural environment are provided in fig.5.

2.5 The SPN graph Model [1,2,32]

A Petri-net model is used here due to its special features [32].a In particular, we use the stochastic Petri-net (SPN) model in a form of a graph and we take the advantage of the SPN properties (timing, parallelism, concurrency, synchronization of events) for our synergistic methodology [1,32]. The graph models described above have the capability of holding structural information about targets. The functional behavior of a target is described by the states in which a particular target could be changed to after an appropriate trigger state is satisfied. A successful and powerful model capable of describing (or modeling) the functional behavior of a system is the Stochastic Petri-net (SPN). Thus, in order to maintain the structural features of the graph model and the functional features of the SPN model, a mapping is presented here, where the SPN model is transformed into a SPN graph model as follows: m : LG → SPN, where, N_i → {P_i}, a graph-node of the L-G graph correspond into a number of SPN places , and {a_{ij}} → {t_{ij}}, relationships corresponds into SPN transitions. In other words we transfer the structural properties of a graph node on to a set of SPN places. This means the different states of the same object correspond to equal number of different SPN places. For instance, a SPN place Pk carries the structural properties of same object at the state k and all structural deformations of that object associated with the state k. There will be cases where the structural properties of an object are the same for all SPN places corresponding to this object. This means the functional behavior of the object changes but the structural properties remain the same. The relationships among graph nodes are represented by transitions on the SPN graph. Thus, the SPN graph will carry functional transitions that fire and transfer the object from the state j to state k and structural transitions that carry structural relationships among the objects structural features. Thus, the SPN graph carries not only the functional properties of the SPN but also the structural properties of the L-G graph as well.

Figure 6 illustrates the SPN graph of an object that has four different states (Places P_i, i=1,2,3,4). Each place P_i has its own structural features transferred from the corresponding graph node N_i. The transitions t_{14} and t_{43} represent relationships among the same parts of a target and a stochastic distribution of time required to fire that transition. The t_{21} transition requires no time to fire.

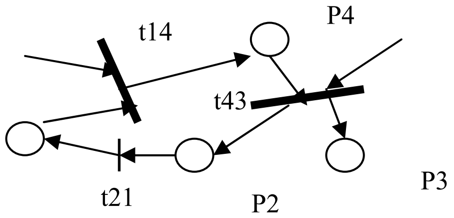


Figure 6. An example SPNG model.

2.6 The Isomorphism of SPN graph onto neural network (NN) for Learning [1,2,29]

The SPN graph model has to deal with a great volume of information due to many functional states associated with each object. Thus, the interpolation and correlation of such amount of information becomes difficult and time consuming. An alternative solution is the use of a neural network model. The neural model alone, however, is not able to efficiently deal with structural representation of knowledge, particularly in a way that can be developed and/or understood by humans. A possible solution proposed here is the transferring of the SPN structural and functional features into a NN via an isomorphism: $I : SPN \rightarrow NN$, where a place P_i is represented by a set of neurons at the NN structure, $P_i \rightarrow \{n_{ik}\}$ $i,k\in Z$, where n_{ik} are a neuron-nodes of a NN, and transitions $t_{ij} \rightarrow w_i$, where w_i represents a threshold logic net). More specifically, each place of the SPN represents structural and a functional state, by assigning a set of neurons to represent the place means the multiple copies of the same structural information are assigned to neurons. Although this isomorphism contains no mapping to a known learning structure, it demonstrates the capability of neural-like models to contain the appropriate information. Note that in this case a multi-layer neural network will be used and a back-propagation algorithm will be selected for the learning process.

2.7 Synergistic Model [1,29]

The general configuration of the synergistic model is presented in figure 7. It shows in a global view the transition of the structural information from the graph to SPN and the functional and structural information of the SPN graph into a neural network. More specifically, the L-G graph provides a powerful description of the image structural features presented in an event, the SPN model offers a way to describe functional behavior of the changes or operations, and the NN model provides the capability of extracting and learning behavioral patterns. The framework described organizes these models in a synergistic hierarchical manner, where the advantages of one model at a particular level are employed by the model of the next level. More specifically, the L-G graph maps its structural descriptions into the SPN model by converting it into a SPN graph that holds both structural and functional representations. Due to high complexity (a great volume of states for each object) at the SPNG level, a search process becomes difficult and time consuming. Thus, by mapping the SPN graph features into a NN model the complexity problem becomes tractable based on the ability of the neural nets to generalize and abstract, providing relevant, though difficult to interpret, models of behavioral categories and predictions. Note that some important features of this synergistic model are the ability of the neural part to learn certain sequences of actions and predict the output before their completion, also the ability to analyze a sequence of actions and determine the possible "causes".

In a more detailed way the synergistic model works as follows. After the segmentation process the model develops the L-G graph by extracting the structural features of the segmented regions and their relationships. Thus, the entire structural information of an image is represented in a form of an L-G graph. The L-G graph is mapped on the SPN structure by generating the SPN graph as shows graphically figure 8.

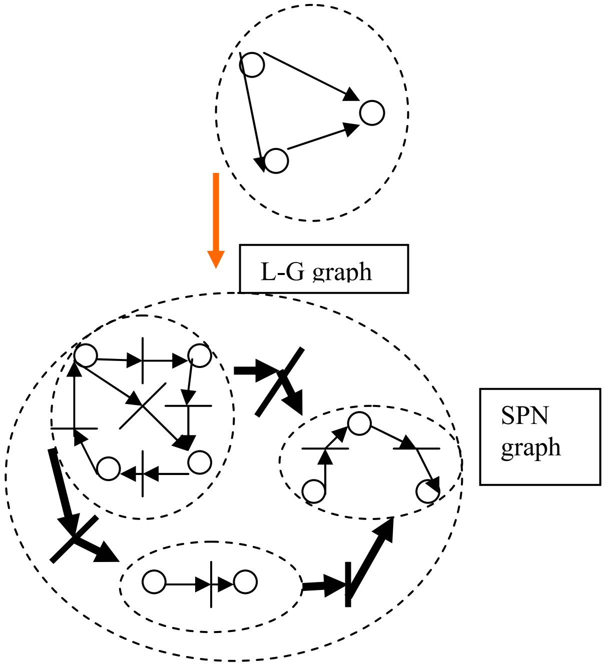


Figure 8: It illustrates the transformation of the LG graph into SNP

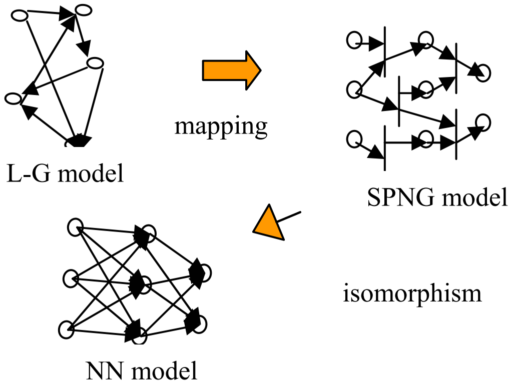


Fig. 7: The general configuration of the synergistic model

In particular, the figure 8 shows a graph with three nodes and their relationships in a form of arrows. The corresponding SPN graph represents these nodes with three sets of places and their functional transitions, while the structural relationships are represented by thick arrow transitions. More specifically, each thick arrow transition represents a set of transitions between set of SPN places as the L-G graph predetermines.

Now, the isomorphism of the SPN graph onto the NN model takes basically place by the transformations illustrated in figure 9. In particular, a place that fires simultaneously to two different places is converted onto a neuron connected directly to two other neurons, also a SPN place that fires to two different places according to a certain probability assigned to each transition is converted into two parallel neurons with the same characteristics but independently connected, finally a single place that fires to another single place is represented by a neuron connected to another neuron.

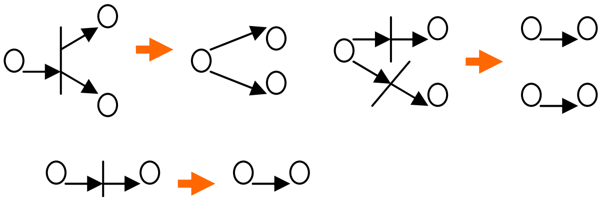


Figure 9: The transformation of SPN basic functional connections into neurons connections

3. REPRESENTATION OF ACTIONS AND EVENTS [29]

So far the methodology discussed above was mainly associated with the information and its various representations in a single image. In this section we provide definitions, descriptions and representations associated with sequences of images or videos and their relationship with the methodology proposed here.

Definition: An action A_m represents the mapping from a state (S(i,t) $\rightarrow$ S(j,t') : A_m : $\Sigma x S \rightarrow S$, where Σ represents the set of actions and S the set of states of a target. Thus, an action A_m could be described as the SPN graph that interrelates the same target into different (or consecutive) frames.

It is known that an event is the result of a sequence of actions, and this event is characterized both by structural and functional information. The recognition of a human activity is strongly related with the ability of describing and interrelating events.

Definition: An event E_i (T_i,T_j) between two targets T_i,and T_j is the result of a set of actions A_m executed by a certain order on these targets.

Figure 10 illustrates an event. Event (catching a ball; target-1 the hand, target-2 the ball), Actions (the five fingers change status from the "open" state into the "closed" state due to coming ball in a synchronized manner described by the SPN).

3.1 Recognition of Actions, Activity and Events

The recognition of an action, based on its definition, is based on the efficient representation of the SPN graphs associated with the states and transitions involved with the particular target. The recognition process is actually a matching of the states, transitions and their order of appearance against the order of target-states and transitions available in a database. The interpretation of an event is actually the recognition of the sequence of actions and activities involved in an event. This means that certain sequences of actions will represent events in a database, thus, when a sequence of actions has been extracted from a video, the sequence goes against the sequences available in the events database. The human activity is defined as a sequence of actions and a set of events associated with these actions. Thus, the recognition of the human activity will be the recognition of the actions and events involved within. Here the system has to be appropriately trained with a variety of activities before its use in real examples.

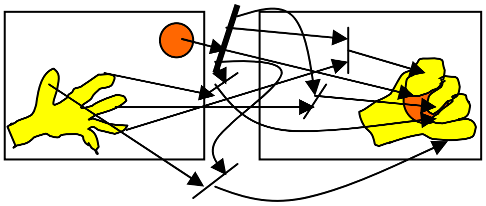


Figure 10. An example of a catching event using an SPN graph for two consecutive frames: A human hand. and a ball : Frame –1: the hand in the open state and a ball in the area coming to its direction; Frame-2 : the hand into a closed state catching the ball. The ball plays the role of the triggering action to make the transition of the targets parts (fingers) from one state into another.

Activity based on L-G Patterns of Behavior: An activity could be also expressed as a sequence of L-G behavioral patterns. This means that at each state an L-G graph can describe the action of an object. Figure 11 illustrates three L-G graphs where their "synthesis" in a form of association will represent an activity. In particular an activity related with a set of objects could be described as the synthesis of the L-G graphs or the SPN graphs (in other words the L-G or SPN patterns) of each object state appeared at each frame and the association of all these L-G or SPN patterns. One has to make a distinction here that an event is a completion of an activity or a set of activities.

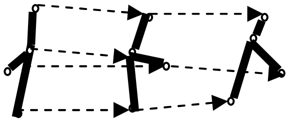


Fig. 11. The connection of three consecutive movements (actions) of a human body based on the L-G graph representation. Note here that the centroids are located on the skeleton of each human standing. These centroids may represent regions of connection points of singularities

4. MULTIPLE CAMERA VIDEO SYSTEM [29]

To implement a system that understands human activity, the viewing system should be able in general to image the tracked human(s) in a broad area over a period of time that is long compared to the events of interest.

Using a fixed single camera has limitations since it restricts tracking to a very narrow area due to the restricted viewing angle of the system. A moving camera with some degree of rotational freedom increases capability in a large environment, but complicates the implementation by adding the motion estimation of both the viewing system and the subject of interest. In the system described here, multiple fixed cameras mounted in the area of interest to track and monitor the motion of individuals in sequences of monocular images are used. Since occlusion is view angle specific, multiple cameras will reduce the chance the occlusion is present in all views. Multiple cameras will also alleviate the difficulty when certain views are confused. To improve the quality of the segmentation and tracking, the system will also use information from multiple calibrated cameras. Although the fusion of information from multiple cameras will improve the segmentation and tracking, there will still be expected levels of error and low confidence observations. To further improve the results, the system to be implemented will not be strictly feed-forward, from low-level operators to high level recognition processes, but will rely feedback from the recognition process and the casting of an activity in larger context.

4.1. Detection of Moving Objects

The system will be initialized by acquiring measurements of the scene over a number of video frames. The goal is to build a powerful representation of the background which will allow us to extract the moving objects in the scene more robustly. Towards this objective, we propose using an eigenspace representation of the background. What is interesting about this representation is that when an image containing new objects is projected onto the background eigenspace, then its reconstructed counterpart does not contain the new objects anymore. This representation needs to be updated over time to account for changes in lighting conditions and new objects in the scene. In a long term, this eigenspace will describe the range background appearances that have been observed. The key to using the eigenspace background representation is on our ability to update it efficiently over time. We will deploy recent results in numerical mathematics (recursive Singular Decomposition techniques) to implement incremental approaches that will allow us to update the eigenspace of the background in real time. To further improve the segmentation results, we will use (i) fusion from segmentations obtained using other viewpoints (multiple cameras), (ii) information from the tracking component (predicted locations), and (iii) feedback from the high-level human activity recognition component.

4.2. Target Extraction and Representation

Connected components analysis, and morphological operations will be used to improve segmentation quality. Each segmented object will then be divided into a number of regions which will be used for building its L-G graph. Specifically, each human will be divided into a number of regions using color, texture, and motion information. Each region will be represented by a mixture of Gaussian learned by using the EM algorithm. In this task the automated detection, extraction and representation in graph forms of a variety of targets, such as human, objects, or other actors, will be performed. In particular, each segmented image frame will be described by the L-G graph model. The L-G graph model provides a flexible representation of the targets and their surrounding in an image frame independent from rotation and shifting. The target detection problem presents two possible scenarios: a) the targets are known to the system, and b) the targets are unknown. For the first scenario, the target is known to the system, it means that its L-G model is available. Thus, the system searches for the target L-G graph form at the image L-G graph representation by using a fuzzy like matching algorithm. When a sub-L-G graph (from the image L-G graph) matches the targets specifications, then that sub-graph is extracted and temporary saved for the next processing task described below. In case that the target is unknown, then the L-G graph of the image frame is exhaustively searched and all possible recognizable or desirable sub-L-G graph forms are extracted and interrelated with their surrounding are saved for further use in later frames.

4.3. Tracking and Association of Moving Objects

A common approach to tracking non-rigid objects is based on using high-level parametric models representing the various object parts (e.g., legs, arms, trunk, head etc in the case of human tracking) and their connections to each other. However, these methods are difficult to apply in real-world scenes due to the difficulty of acquiring and tracking the requisite model parts (e.g., specific joints such as knees, elbows or ankles in human tracking). Another problem with this approach is that a separate model is required for each type of objects to be tracked (e.g., humans, animals, etc.). In this project, we propose to build dynamic models of appearance of the objects being tracked. These models will enable robust tracking but at the same time provide useful information for classifying moving objects as rigid or non-rigid. The key to building dynamic models of appearance lies on using an incremental eigenspace approach such as the one described earlier for background modeling. Using multiple cameras will facilitate this since different views of the object will be obtained quickly. To build the eigenspace of an object, the segmented moving object will be resampled in a canonical frame throughout tracking, despite changes in scale and position. Its estimated bounding box will be used to resize and resample it into a canonical view. To classify the object as rigid vs. non-rigid, the distances of the views used to update the eigenspace will be analyzed (non-rigid objects are expected to create widely different views, thus, the distances will be large compared to the ones obtained when building the eigenspace of rigid objects). To build the larger object graph, the system searches for the sub-L-G graph form of the same target.

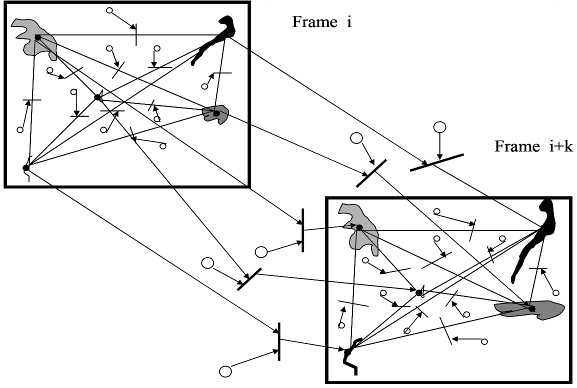


Figure 12: SPN associations of the changes occurred in two image frames. These changes may represent also motion, or moving targets. SPN has timing of occurrence of the changes due to its transitions.

When the sub-L-G graph is discovered, the two graphs are connected by maintaining the relationships between the sub-graphs of the same in consecutive frames. The same process is repeated and the output is the tracking path of a target in a sequence of video frames. When a target is detected in a specific frame, the L-G graph provides the capability to interrelate each recognizable target with any other target in the same frame. The stochastic Petri-net graphs method has the potential to correlate and interrelate changes occurred on images or sequences of images and produce interpretations or discover new knowledge. In order to visually demonstrate the functionality of the SPN method, an illustrative example is presented below, see figure 12.

5. ILLUSTRATIVE EAXPLES:

Associating & Tracking Changes and Moving Objects in Sequences of Images [36]

In this section we provide an illustrative example to present the potential of the methodology described here. In particular, using the SPN Graph by associating and tracking multiple moving targets. In this case we firstly apply a spatio-temporal anisotropic diffusion method that uses the information of the current, previous and next frames in the sequence, see fig.13.

This method is based on the anisotropic diffusion by inserting a temporal variable in the heat diffusion equation. This smoothes out adaptively the areas of spatial and temporal homogeneity. Then a segmentation algorithm is applied on the diffused image. This will produce higher segmentation detail in the areas that are not spatio-temporal homogeneous, see figure 14a.

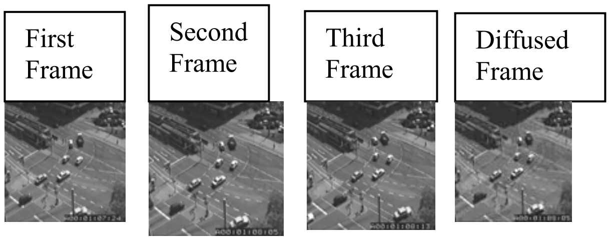


Fig 13: It shows 3 frames and the diffused fame

The Displaced Frame Difference is estimated between the Diffused frame and the current frame, see figure 14b. A process to identify the active regions follows. First, the level of activity is calculated as the ratio of the sum of dfd pixels in a watershed region over it's area. A thresholding operation follows to detect the most active areas in the frame which are also decomposed into watershed regions. Figure 14c shows the areas that represent differences tracked from frame to frame.

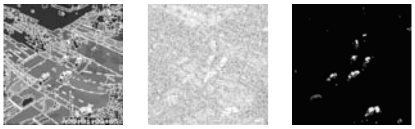


Figure 14:a) segmented image, b) displaced frame difference, c) representation of areas of differences tracked from frame to frame

SPNG Associations: Here we presents the results of SPN graphs associations for detecting formations from moving targets or patterns, see figure 15. More specifically, when the changes are detected and tracked in different frames, the Local Global (L-G) graph method is used to establish a local graph for each region. These region-graphs then associated for developing the global graph that associated all the region-graphs. This is the association pattern that represents the formation. By tracking these formations we have better understanding of the changes that take place in sequences of frames.

Tracking Patterns of Formations: Here, we present the SPN graph formations. In particular, in each frame the changes based on motion are detected and extracted and their shapes are isolated from the background image. Then, these shapes, that may represent moving targets or objects, are described by using Local graphs and their relative locations in the frame is associated with Global graphs, figure 16 &17. This means that these changes (objects or targets) are fully represented by the L-G graphs. At this point, we take these L-G graphs representations (or formations) from each frame and we again associate them with SPN graphs in order to explain (or represent) their transitions from one frame to the next (or from one state to the next). As it is known well SPN is capable for representing state transitions efficiently [29,36].

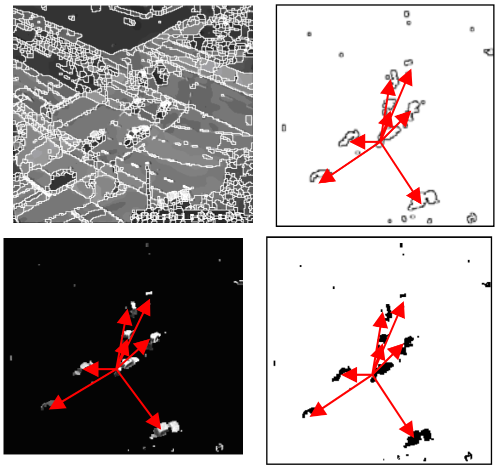


Figure 15: It shows the SPN graph associations (in red color) for formations (frame No-4)

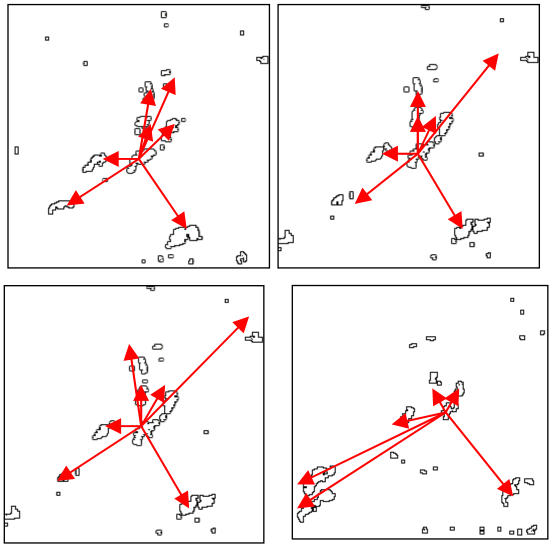


Figure 16: Formations (in red color) taken from different moving patterns of targets in different frames

6. CONCLUSION

The paper presented a synergistic methodology for associating multiple patterns of actions that take place in events by using a video. The methodology is composed by several techniques, such as image segmentation, region growing, local-global graphs, stochastic Petri-nets and neural nets. Each of these techniques contributes to this synergy in a way by offering their advantages for accomplishing the goals, which are extracting rules and patterns and learning behavior in multiple events automatically. This work is a part of the project HMI theory.

Acknowledgement

This work is partially supported by AIIS Inc., AFRL and NSF grants.

References

[01] N. Bourbakis, A Neural-based KB using SPNGs in sequence of images, AIRFORCE & SUNY-B-TR-1991,45 pages, Nov. 1991.

[02] J. Gattiker& N.Bourbakis Representation of Structural and Functional Knowledge using SPN Graphs, Proc. IEEE Conf. on SEKE 1995, MD.

[03]N. Bourbakis, A survey on tracking human activities and body patterns, AIIS-ITRI-TR-2006

[04] J.Yamato, J.Ohya and Kenichiro, "Recognizing human action in time-sequential images using Hidden Markov Model", IEEE Conf CV. 1992, pp.379-385

[05] C.Bregler, "Probabilistic Recognition of Human Actions", UCB-TR-May-1996,p.28

[06] C.Bregler, "Learning and Recognizing Human Dynamics in video sequences" IEEE Conf. on CVPR, Puerto Rico, June 1997

[07] M.Yeasin and S.Chaudhuri, "Visual understanding of dynamic hand gestures", PR Journal 33,11,2000, 1805-1817.

[08]N.Bourbakis,"Extraction, Tracking And Recognition of Targets In Sequences Of Images", *Int. Journal AIT*, vol. 11, no. 4, 2002.

[09] P.Yaun, A.Mogzadeh, D.Goldman and N.Bourbakis A fuzzy-like approach to edge detection in colored images, IAPR Pattern Analysis and Applications, vol.4,4,272-282,2001

[10] N.Bourbakis, Emulating human visual perception for measuring differences in images using an SPN graph approach, IEEE T-SMC, 32,2,191-201, 2002

[11]N.Bourbakis, P.Yuan and P.Kakumanu, Representing and recognizing 3-D objects in images using LG, IEEE T-SMC 2007

[12] McAllister, G.; McKenna, S.J.; Ricketts, I.W. Hand tracking for behavior understanding Image and Vision Computing Volume: 20, Issue: 12 October 1, 2002, pp. 827-840

[13] Rittscher, J.; Blake, A.; Roberts, S.J. Towards the automatic analysis of complex human body motions Image and Vision Computing Volume: 20, Issue: 12 October 1, 2002, pp. 905-916

[14] Gong, Shaogang; Ng, Jeffrey; Sherrah, Jamie On the semantics of visual behaviour, structured events and trajectories of human action Image and Vision Computing Vol: 20, 12 October 1, 2002, pp. 873-888

[15] Stauffer, C. Grimson, W.E.L. Learning patterns of activity using real-time tracking IEEE Transactions on Pattern Analysis and Machine Intelligence, Volume: 22, Issue: 8, Aug 2000, pp.747-757

[16] Haritaoglu, I.; Harwood, D.; Davis, L.S., W4: Real-time surveillance of people and their activities IEEE Trans. Pattern Analysis and Machine Intelligence, 22, 8, Aug 2000, pp. 809-830

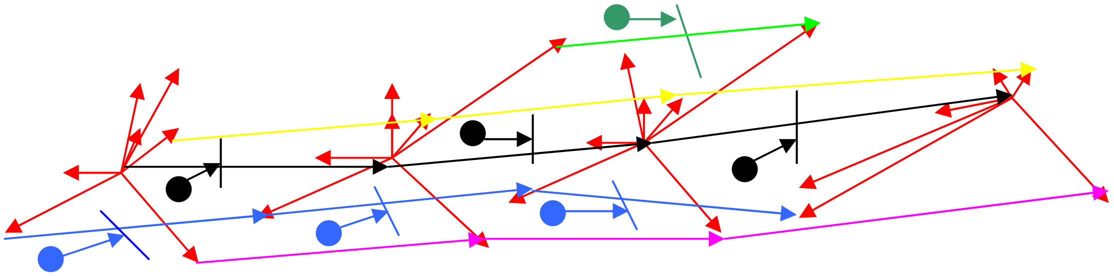


Fig. 17: Tracking patterns of Formations in different frames. Due to limitations of colors only 4 pattern formations are illustrated. The SPN graph represents the transition from one state to the next. The circles represent the tokens that activate the transition. Due to the complexity of the diagram only a few tokens and transitions are presented in this picture. Color lines represent the same regions (targets) moving in different locations in time. It shows also the formations (associations) moving in time.

[17] Ivanov, Y.A.; Bobick, A.F., Recognition of visual activities and interactions by stochastic parsing, IEEE Trans. Pattern Analysis and Machine Intelligence, 22, 8, Aug 2000, pp. 852-872

[18] Junji Yamato, Jun Ohya, Kenichiro Ishii: "Recognizing Human Action in Time-Sequential Images using Hidden Markov Model", Proc. of IEEE Conference on Computer Vision and Pattern Recognition 1992

[19] J. Yang, Xu Yangsheng and Chiou S. Chen, "Human Action Learning via Hidden Markov Model", IEEE T-SMC- Part A: vol. 27, January (1997), 34--44.

[20]A.Moghddamzadeh and N. Bourbakis, "Segmentation of Color Images With Highlights and Shadows Using Fuzzy Reasoning", *SPIE Conf. Electronic Imaging*, pp.300-310, 1995.

[21] A. Moghaddamzadeh and N. G. Bourbakis, "A Fuzzy Region Growing Approach for Segmentation of Color Images", *PR Society Journal of Pattern Recognition*, vol. 30, no. 6, pp. 867-881, 1997.

[22] G. J. Klinker, S. Shafer and T. Kanade, "Image Segmentation And Reflection Analysis Through Color", *Image understanding workshop*, San Matteo, California, pp. 838-853, 1988

[23] G. J. Klinker, "A Physical Approach To Color Image Understanding", *A. K. Peters*, Wellesley, Massachusetts, 1992.

[24] Y. Xiaohan and J. Yla-jaaski, "Image Segmentation Combining Region Growing And Edge Detection", *11th Int. Conf. on Pattern Recognition*, The Netherlands, 1992

[25] Narendra Ahuja, Byong An and Bruce Schachter, "Image Representation Using Voronoi Tessellation", *Computer Vision, Graphics and Image Processing*, vol. 29, pp. 286-295, 1985.

[26] Narendra Ahuja, "Dot Pattern Processing Using Voronoi Neighborhoods", *IEEE T-PAMI,* vol. 4, no. 3, pp. 336-342, 1982.

[27] E. Kubicka, G. Kubicki and I. Vakalis, "Using Graph Distance In Object Recognition", *1990 ACM Eighteenth Annual Computer Science Conference Proceedings, ACM, New York, NY*, pp. 43-48, 1990

[28] N. Bourbakis, "A Rule-Based Scheme For Synthesis Of Texture Images", *Int. IEEE Conf. on Systems, Man and Cybernetics*, Fairfax, VA, pp. 999-1003, 1987.

[29] N.Bourbakis, J.Gattiker and G.Bebis, Representing and interpreting human activity and events from video, IJAIT, vol.12.1.2003

[30]Y. Yacoob & M. J. Black, "Parameterized Modeling and Recognition of Activities", JCVIU, vol. 73, no. 2, pp. 232--247, 1999.

[31] Douglas Ayers and Mubarak Shah. "Monitoring Human Behavior in an Office Environment", Workshop on Applications of Computer Vision 1998

[32] T.Murata, *Petri nets: properties, analysis and applications*, Proc. of the IEEE,7,4,1989

[33] H-G Kang and D.Kim, Real-time multiple people tracking using competitive condensation, Pattern Recognition Journal, vol.38, 7, pp. 1045-1058, 2005.

[34] C. Lerdsudwichai et.al., Tracking people with recovery from partial and total occlusion, Pattern Recognition Journal, vol.38,7, pp. 1059-1070, 2005.

[35] P.Kakumanu, et. al., Image chromatic adaptation using ANNs for skin color adjustment, IEEE Conf. TAI-04, FL, Nov. 15-17, 2004, pp. 478-485.

[36] N. Bourbakis, Human-Machine Interaction (HMI) Theory: Multi-registration for motion analysis of multiple humans, WSU-ITRI-TR-2005.

[37] Brand, M.; Kettnaker, V., Discovery and segmentation of activities in video IEEE Transactions on Pattern Analysis and Machine Intelligence, 22, 8, Aug 2000, pp. 844-851

[38] Oliver, N.M.; Rosario, B.; Pentland, A.P., A Bayesian computer vision system for modeling human interactions IEEE Transactions on Pattern Analysis and Machine Intelligence, 22, 8, Aug 2000, pp.831-843

[39] D.M. Gavrila, "The Visual Analysis of Human Movement: A Survey," *Computer Vision and Image Understanding*, vol. 73, No. 1, 1999.